From symbols to icons: the return of resemblance in the cognitive neuroscience revolution

Article (Published Version)

Synthese (2018) 195:1941–1967
https://doi.org/10.1007/s11229-017-1578-6

S.I. : NEUROSCIENCE AND ITS PHILOSOPHY

From symbols to icons: the return of resemblance in the cognitive neuroscience revolution

Daniel Williams[1] · **Lincoln Colling**[2]

Received: 1 June 2017 / Accepted: 21 September 2017 / Published online: 13 October 2017

Abstract We argue that one important aspect of the "cognitive neuroscience revolution" identified by Boone and Piccinini (Synthese 193(5):1509–1534. doi:10.1007/s11229-015-0783-4, 2015) is a dramatic shift away from thinking of cognitive representations as arbitrary symbols towards thinking of them as icons that replicate structural characteristics of their targets. We argue that this shift has been driven both "from below" and "from above"—that is, from a greater appreciation of what mechanistic explanation of information-processing systems involves ("from below"), and from a greater appreciation of the problems solved by bio-cognitive systems, chiefly regulation and prediction ("from above"). We illustrate these arguments by reference to examples from cognitive neuroscience, principally representational similarity analysis and the emergence of (predictive) dynamical models as a central postulate in neurocognitive research.

Keywords Cognitive neuroscience revolution · Cognitive representation · Structural resemblance · Predictive processing · Representational similarity analysis · Emulation · Bayesian brain · Bayesian networks · Iconic · Mechanistic explanation · Information-processing · Free energy principle · Homeostasis · Predictive brain

✉ Daniel Williams
dw473@cam.ac.uk

Lincoln Colling
lincoln@colling.net.nz

1 Faculty of Philosophy, Trinity Hall, University of Cambridge, Cambridge, UK

2 Department of Psychology, Darwin College, University of Cambridge, Cambridge, UK

1 Introduction

Boone and Piccinini (2015) have recently argued that cognitive neuroscience constitutes a revolutionary break from traditional cognitive science, distinguished by its abandonment of the autonomy of psychology from neuroscience in favour of a multilevel mechanistic approach to neurocognitive explanation. In this paper we identify one important aspect of this revolution: a dramatic shift away from thinking of cognitive representations as *arbitrary symbols* towards thinking of them as *icons* that replicate structural characteristics of their targets. This shift has received increasing attention in the philosophical literature in recent years (e.g., Churchland 2012; Cummins 1989; Grush 2004; Gładziejewski and Miłkowski 2017; O'Brien and Opie 2015b; Ryder 2004; Williams 2017). We aim to clarify what it consists in, and explain why it has occurred.

To this end, we identify two driving forces behind this transition to resemblance-based accounts of cognitive representation. The first comes from a better appreciation of what mechanistic explanation of information-processing involves. We argue that broadly *a priori* arguments for thinking such mechanistic explanations mandate resemblance-based representational architectures (as advanced by, e.g., O'Brien and Opie 2004, 2015b) are vindicated by an examination of the nature of such explanations in contemporary cognitive neuroscience. The second comes from a better appreciation of the *problems* solved by nervous systems, especially as exemplified in a recent explosion of fruitful work that focuses on *regulation* and *prediction*. We illustrate these lessons by drawing on "predictive processing" (Clark 2016; Friston 2009; Hohwy 2013) and its conception of the mammalian cortex as a general-purpose model-making machine.

We structure the paper as follows. In Sect. 2 we outline both the nature of and the motivation for the symbolic approach to cognitive representations in traditional cognitive science. In Sect. 3 we present an overview of the resemblance-based approach to cognitive representation, and distinguish this iconic account from both "detector"-based accounts and anti-representationalist research programmes. We then identify the two chief factors that have driven cognitive neuroscience towards resemblance-based representations: the implications of a mechanistic approach for the explanation of information-processing systems (Sect. 4), and a greater appreciation of the *problems* solved by bio-cognitive systems (Sect. 5). We conclude in Sect. 6 by identifying the chief implications of and challenges for an iconic cognitive neuroscience.

2 Symbols and syntax in traditional cognitive science

Boone and Piccinini (2015) identify three central features of "traditional cognitive science" (henceforth TCS), the research programme they take cognitive neuroscience to have superseded. First, TCS understands cognition as a form of *digital computation*, as exemplified in Newell and Simon's (1976) famous "physical symbol systems" hypothesis, according to which the essence of intelligence lies in the rule-governed manipulation of symbolic data structures. The second was closely allied to this commitment: cognitive-scientific explanation is *functional* explanation that abstracts away

from "implementational" details in the brain, in much the same way that a software description of processes implemented in commercial digital computers abstracts away from details of their hardware. Third, this led to what Boone and Piccinini (2015, p. 1510) call the "strong autonomy assumption": scientific psychology is theoretically independent from neuroscience; the former concerns the functional architecture of cognition, the latter merely the contingent matter of implementation.

As Boone and Piccinini (2015, p. 1510) note, this assemblage of commitments has the embarrassing upshot of making "cognitive neuroscience" sound like an oxymoron—embarrassing, because cognitive neuroscience has emerged as the "mainstream approach to studying cognition." This new research paradigm, they argue, abandons the strong autonomy assumption at the core of TCS in favour of a *multilevel mechanistic approach* to neurocognitive explanation: a theoretical "revolution [that] requires a different way of thinking about levels, cognitive explanation, representation, and computation" (Boone and Piccinini 2015, p. 1513)

We are in full agreement with Boone and Piccinini's thesis, and return in more detail in Sect. 4 to the nature of multilevel mechanistic explanation. Because our focus here is on that part of the revolution they identify that concerns the nature of representation, however, we start by getting clearer about how this construct was initially understood in TCS.

As the reference to Newell and Simon's hypothesis indicates, the core idea is this: cognitive representations are *symbols*. It is notoriously difficult to specify what is distinctive about *symbolic* forms of representation (cf. Marcus 2003). With respect to *public symbols*, at least, an influential tradition differentiates them as *arbitrary* representations whose semantics is owed not to properties of the symbol *itself* but rather to convention and interpretation (e.g. Peirce 1931–1958). When TCS first emerged, a similar idea prevailed: symbols are *formal entities* individuated by their *syntactic* properties in a manner familiar from formal logic, such that their systemic role within a broader mechanism—a physical symbol system—is a function of their syntax *rather than semantics* (cf. Haugeland 1989). This characteristic of digital computers lies behind their popular characterisation as "syntactic engines" (Dennett 2013; see below).

Given this, the idea of explaining *intelligence* in terms of symbol manipulation has seemed hopeless to many (e.g. Searle 1980). What *we* do, after all, at least seems to be a function of *what* we think and want—that is, the *contents* of our mental states. Nevertheless, there were at least three important ideas that made the vision of a symbolic mind dominate TCS.

First, ground-breaking work in formal logic at the turn of the twentieth century pioneered by figures like Boole, Frege and Russell demonstrated that rules of reasoning and thus *semantic relations* can be mimicked (up to well-known limitations) by purely *syntactic* or *formal* operations on symbols—that is, by purely algorithmic operations that are insensitive to the contents of the expressions they are defined over.

Second, the work of Turing (1937) and others demonstrated that relatively simple physical mechanisms can be designed to function as general purposes syntax-sensitive inference machines. To many, this combination of ideas strongly suggested a vision of the *brain* as such a machine: a general-purpose syntax-sensitive inference engine of the sort envisaged by Newell and Simon (1976). This vision was bolstered by pioneering

work by McCulloch and Pitts (1943) that revealed how simple neuronlike nodes could compute logical functions.

Finally, a widely-held conviction with deep roots in the philosophical tradition and still held by many today (e.g. Fodor and Pylyshyn 2015; Gallistel and King 2011) is that certain conspicuous features of human cognition can only be explained by recourse to operations on a symbol system with a combinatorial syntax and semantics. For example, Chomsky (1959) famously argued that the generative flexibility exhibited in language comprehension and production requires a compositionally structured symbol system, an insight that was extended to many other domains where systematic and productive representational capacities seem—*prima facie*, at least—to be required (Fodor 1975). Digital computers are physical machines capable of exhibiting such characteristics. Therefore—the argument goes—we are digital computers.

The upshot of these three considerations was a conception of cognitive representations as *formal* (syntactically individuated) entities amenable to syntax-sensitive transformations. Before turning to explain in what way this conception has been largely supplanted in cognitive neuroscience, we clarify an important issue.

It is sometimes suggested that "syntax-sensitive" processing—that is, the idea that an information-processing system is sensitive only to the syntactic properties of its representational vehicles, and *not* their semantic properties—is a characteristic of *any* information-processing system, not just symbol-based ones. Dennett (2013, p. 178), for example, writes:

> Brains… are supposed to be *semantic engines*. What brains are *made of* is kazillions of molecular pieces that interact according to the strict laws of chemistry and physics…[B]rains, in other words, are in fact only *syntactic engines*… *A genuine semantic engine*, responding *directly* to meanings, is like a perpetual motion machine—physically impossible. So how can brains accomplish their appointed task? By being syntactic engines that *track* or *mimic* the competence of the impossible semantic engine (emphasis in original).

This widespread view conflates two different meanings of "syntax," however. On one reading, "syntax" refers to the individuation of *formal symbols*, and is itself an autonomous level of description that is multiply realisable at the physical level (Fodor 1975). On another reading, "syntax" just refers to those physical properties by which representational vehicles can perform a systemic role within a mechanism. Of course, *in this latter sense* all information-processing mechanisms are "syntactic engines." The question is what differentiates systems that are *merely* sensitive to the physical properties of their component parts from genuinely *information-processing* systems—that is, systems whose behaviour is a function of the *representational properties* of their component parts?

An influential proposal associated with TCS (see Sect. 4 below) is that such information-processing systems are carefully crafted mechanisms whose behaviour is sensitive to the *syntactic* properties of their parts *in the former sense*. These are systems in which—as Haugeland (1989, p. 106) famously put it—"if you take care of the syntax, the semantics takes care of itself."

Is there an alternative?

3 Iconic representations

The history of philosophy has given rise to two major rival views about the nature of mental representation (cf. Waskan 2006). On one view—the intellectual provenance of TCS—mental representations are language-like entities and cognition is a matter of what Hobbes famously called *ratiocination*. In contrast to this, an older tradition—associated with figures like Aristotle, the Scholastics, and the British empiricists—holds mental representation to be founded on *similarity* to the mind's objects, and takes as its paradigms of representation things like pictures, maps, diagrams, and scale models (O'Brien and Opie 2015b). In Peirce's (1931–1958) vocabulary, such representations are *iconic*, not *symbolic*: they represent by *resemblance*. In addition to the introspective vindication of this approach, an influential tradition in philosophy contends that only the similarity between the mind's representations and the world could explain our ability to use the former to reason about the latter (cf. Cummins 1989; Isaac 2013; see Sect. 4 below)

Iconic approaches to mental representation fell on hard times throughout most of the twentieth century. After all, once we must situate mental representations inside the brain, it just seems obvious that they cannot literally reduplicate the world and its heterogeneous contents (Cummins 1989, p.31).

We contend that this is *not* obvious. In fact, we contend that cognitive neuroscience positively vindicates this ancient view about internal representation. Before turning to justify this claim, however, we first clarify what is meant by iconic representation, and distinguish our view from two other prominent ideas about representation in cognitive neuroscience.

3.1 Resemblance

First, the relevant kind of resemblance for brain-bound cognitive representations is "second-order structural resemblance" (O'Brien and Opie 2004). Roughly, this resemblance relation obtains between two domains when the pattern of relations among the elements in one recapitulates the pattern of relations among the elements in the other.[1] This kind of resemblance relation is thus straightforwardly consistent with a vapid physicalism: it demands only that neural structures, properties, and processes in the brain can instantiate the relevant structure of the domains they represent. Further, it explains how iconic representations can capture the more rarefied characteristics of what we think about. Insofar as such phenomena are individuated by their role in broader structures—spatial structures, categorical structures, dynamical or causal structures, and so on—iconic representation just requires neural vehicles to replicate that structure. For this reason, iconic representations are often called "structural" or "S-representations" (Cummins 1989; Gładziejewski and Miłkowski 2017; Ramsey 2007).

Second, second-order structural resemblance is weaker than isomorphism or homomorphism (O'Brien and Opie 2004). This is familiar from public iconic representations

[1] See O'Brien and Opie (2004) for a more technical analysis.

such as cartographic maps. Such iconic representations are often heavily idealised, selective, and sometimes purposefully distortive (Horst 2016). For this reason, resemblance is a profoundly *graded notion*. Consequently, representational *value* is similarly graded in iconic representations. Maps are not true or false, but *better* or *worse*, where this continuous value is a function of their practical utility. Intuitively, you and I might both have pretty accurate maps of London, but mine might be better than yours, Sally's might be better than both of ours, and Bob might have a map *too* accurate and detailed to be of any use in getting around. As for maps, so for iconic representations more generally.

Third, philosophers are fond of pointing out that resemblance is not sufficient for representation (Crane 2003; Goodman 1969). But *no* dyadic relation is sufficient for representation. We follow Peirce and a long tradition of subsequent authors in assuming that representation is essentially a *triadic* relation that exists between: (1) a representational vehicle or set of vehicles; (2) the target of the representation; and (3) the system that *uses* or *interprets* the former to coordinate its behaviour with the latter (O'Brien 2015a; cf. O'Brien and Opie 2004). The claim that cognitive representations are *iconic* is thus the claim that what explains the brain's ability to *exploit* its internal states as *representations* of target domains is that they *resemble* such domains (Shea 2014). Importantly, this triadicity undermines concerns about the reflexivity, symmetry, and ubiquity of resemblance relations:[2] my map might structurally resemble itself and an infinite number of other things (which in turn resemble my map), but it is only the similarity between its internal structure and London that I exploit in making my way to King's Cross (O'Brien 2015a). In a slogan: "no representation without exploitation."

We join a growing chorus of voices in advancing the thesis that cognitive neuroscience commits one to an understanding of cognitive representation as iconic (Churchland 2012; Gładziejewski and Miłkowski 2017; O'Brien and Opie 2015b; Ryder 2004). Before we explain *why* this shift has occurred—and identify evidence for it—we first distinguish our view from two other prominent ideas in the literature.

3.2 Detectors

One prominent view among philosophers is that the leading concept of representation in cognitive neuroscience is the *detector*, a component of the nervous system that responds selectively to (i.e. *detects* or *indicates*) the presence of some environmental stimulus (e.g. Ramsey 2007). Familiar examples are edge detectors in V1 or neuronal populations that respond differentially to the presence of specific faces in one's fusiform "face area." On this view, cognitive neuroscience has abandoned an understanding of cognitive representations as formal symbols in favour of an understanding of nervous systems as complex measuring instruments: intricately structured networks of cells designed to reliably indicate the presence of functionally significant environmental stimuli.

This cannot be right, however.

[2] See Goodman (1969).

First, detection *itself* is not a representational relation (Ramsey 2007). Specifically, detection exists whenever there is a covariational relation between the states of two systems—a relationship that is *ubiquitous* across all physical (cognitive, biological, and even non-organic) systems. As such, detection is most often relevant as a proposal about *how* representations—namely, vehicles *already* supposed to function as representations—*acquire their contents*. Specifically, because symbols evidently do not *resemble* their targets, an influential proposal is that *symbols* acquire semantic properties through the indication relationships they stand in to what they represent—the basis of various kinds of "indicator semantics" (Dretske 1981; Fodor 1987). The claim that *detection* is the primary relation that nervous systems bear to their environments is thus either a call to abandon internal representations altogether (Hutto and Myin 2013; Ramsey 2007; see below), or a proposal about how *symbols* acquire their contents (e.g. Fodor 1987).

Second, a central assumption of cognitive neuroscience is that computations over internal representations *causally explain* intelligence (Bechtel 2008; Boone and Piccinini 2015). But detection *as such* cannot explain intelligence: the kind of exquisite informational sensitivity to environmental contingencies that advocates of indicator semantics appeal to in order to explain representation is an instance of the very intelligence supposed to be explained *by* representation (Churchland 2012; Cummins 1996). As Churchland (2012, p. 97) puts it, "if you are to have any hope of recognizing and locating your place and your situation within a complex environment, you had better know, beforehand, a good deal about the structure and organization of that environment."

Third, and relatedly, many alleged "detectors" are better thought of as components of larger models that function as iconic representations (see Sects. 4 and 5). Place cells in the rat's hippocampus, for example, reliably indicate the rat's presence at a certain location of the environment (Shea 2014). This fact does not *explain* their representational role, however; their functioning as representations is explained by their participation in a larger network of cells comprising "cognitive maps" whose internal structure—roughly, the pattern of coactivation relations in the network—recapitulates the spatial structure of the ambient environment (Gładziejewski and Miłkowski 2017; Shea 2014).

Finally, it is plausible that even extremely simple instances of alleged "detector"-based representations in fact represent by structural resemblance (Morgan 2014). Consider a thermostat, for example—seemingly a *paradigmatic* detection-based system. Nevertheless, what *explains* the thermostat's ability to regulate room temperature is that the pattern of relations among its bimetallic strip curvatures replicates the pattern of relations among ambient air temperatures (see Sect. 4 below) (O'Brien 2015a). This suggests that the core difference between paradigmatic "detector" representations and iconic representations in cognitive neuroscience is that the relevant resemblance relation in the former case is mediated by a simple causal relation, and is purely reactive (Gładziejewski and Miłkowski 2017). As we argue below, the core kinds of cognitive representations in cognitive neuroscience do not have these characteristics. Biological agents are autonomous, endogenously active—not merely *reactive*—systems (Bechtel 2008).

For these reasons, we think an understanding of internal representations as *detectors* is untenable. Nevertheless, Ramsey *is* evidently right that the detector-notion of representation is widely used in cognitive neuroscience. How can we reconcile this fact with our contention that cognitive neuroscience vindicates an iconic approach to representation?

In three ways. First, as noted above, claims about detection are often heuristically essential in constructing and evaluating theories about the broader internal models (i.e. iconic representations) that underpin intelligence (Bechtel 2014; Kiefer and Hohwy 2017). Indication relations are thus often crucial components of the *explananda* that iconic representations are posited to explain. Second, and relatedly, our claim is emphatically *not* that detection is irrelevant to the *explanation* of intelligence. Evidently any adaptive system must be differentially responsive to environmental conditions, and indication relations are thus crucial for the proper *functioning* of iconic representations—for *updating them*, for example (Gładziejewski and Miłkowski 2017). For these reasons, our thesis is consistent with the prevalent role that detection plays in representational theorizing. Specifically, this prevalence should not be confused for the claim that representation *is* detection.

Finally, our thesis must be partially normative. Whilst we argue that iconic representations evidently *are* central to contemporary research in cognitive neuroscience (see below), we acknowledge that a detector-notion of representation is also prevalent in that research.[3] Thus our thesis should be understood as stating that our *best* and most *fruitful* research into the nature of intelligence and adaptive response involves iconic representations. For the reasons just stated and outlined in Sects. 4 and 5 and below, we contend that any research that relies *purely* on detector relations is itself hopeless to explain intelligence.

3.3 Anti-representationalism

Another prominent view is that the cognitive sciences are not just moving away from symbolic representations but from the very idea that internal representations form a core part of the explanation of intelligence and adaptive behaviour (e.g. Anderson 2014; Chemero 2009). On one manifestation of this view, the idea is just that traditional symbol-centred cognitive science paid insufficient attention to the importance of variables like *time*, *action*, and *non-neural* parts and operations in cognitive mechanisms (e.g. Clark 1997). Context-invariant, discrete linguaformal symbols and sequential algorithms, the argument goes, are ill-suited to many basic forms of adaptive success (see Sect. 5) in a way that TCS was oblivious to. These lessons are salutary, and consistent with the ideas we defend here. On a more "radical" manifestation of this stance, however, this hostility to internal symbols is extended to internal representations more generally, which are marginalised as either "peripheral or emergent" (Anderson 2014, p. 162) or eliminated altogether (Chemero 2009) as a result.

We cannot fully address this scepticism here. Nevertheless, we briefly note three things.

[3] We thank an anonymous reviewer for pressing this point.

First, as a claim about how cognitive neuroscience is actually practiced—the focus of this paper—this view is mistaken. Anti-representationalism is profoundly revisionary. Open any mainstream textbook in the discipline and see. As far as we can tell, this radicalism has not borne fruit when it comes to explaining even moderately sophisticated instances of intelligence and adaptive success, such as a rat's ability to navigate a maze (Bechtel 2008). Further, we suspect there are principled reasons why it will fail (see Sect. 5 below).

Second, and relatedly, advocates of radical anti-representationalism (particularly dynamical systems theory approaches) often reject mechanistic explanation in favour of covering law explanations (see Sect. 4.1) (see, e.g., Chemero 2001, 2009). The deficiencies of such explanations in the special sciences are well-known, so we will not recapitulate them here (see Colling and Williamson 2014a; Kaplan and Bechtel 2011a; Kaplan and Craver 2011b). In the current context, however, note that embracing this covering law model of explanation would in effect amount to an abandonment of cognitive neuroscience as Boone and Piccinini (we think rightly) characterise it.

Finally, as we will see, many of these broadly anti-representationalist *concerns* with TCS can be accommodated within a representationalist framework that favours *iconic* representations. Indeed, there is an important irony here. Advocates of contemporary anti-representationalism typically propose that we should approach the processes responsible for intelligence and adaptive behaviour with the resources of dynamical models *instead of* orthodox concepts like *representation* and *computation* (e.g. Chemero 2009). An exciting body of work in contemporary cognitive neuroscience, however, holds that the *brain itself* instantiates dynamical models—that is, that the brain makes use of the very kind of model that advocates of anti-representationalist approaches claim *we* should use to model the brain's interactions with its environment. Such neurally instantiated dynamical models replicate the covariational relations among functionally significant variables, and are iconic representations *par excellence*. Or so we will argue (Sects. 4 and 5). Thus the influence of dynamical systems theory on cognitive neuroscience might turn out to be *exactly the opposite* of what its proponents intended (cf. Bechtel 1998).

3.4 Summary

With these preliminaries in hand, we turn next to justify our claim that cognitive neuroscience embraces an iconic account of cognitive representations. We argue that the shift has been driven from two directions: "from below" (Sect. 4) and "from above" (Sect. 5)—that is, from a greater appreciation of what mechanistic explanation of information-processing systems involves ("from below"), and from a greater appreciation of the *problems* solved by cognitive systems ("from above").

4 The return of resemblance

The core aspect of the cognitive neuroscience revolution identified by Boone and Piccinini (2015) is the shift from purely functional analyses of cognitive systems to multilevel mechanistic explanations of the capacities they exhibit. In this section we

argue that this shift mandates an iconic understanding of cognitive representation. This can be seen in two ways: first, by looking at what mechanistic explanation of distinctively information-processing systems involves—a broadly *a priori* approach (Sect. 4.1); and second, by looking at actual examples of mechanistic explanation of information-processing in cognitive neuroscience (Sects. 4.2, 4.3).

4.1 Mechanistic cognitive science entails iconic representation

First, then, sciences such as biology and cognitive neuroscience adopt a mechanistic approach to scientific explanation (Bechtel 2008; Craver 2007). This approach is distinct from the dominant philosophical model of scientific explanation assumed throughout much of the 20th century—namely, the "covering law" or "deductive-nomological model"—according to which the primary goal of explanation is to uncover universal or statistical laws and then *derive* statements of phenomena from statements of such laws and initial conditions (Hempel and Oppenheim 1948, 1953). Bechtel and Abrahamsen (2005, p. 3) offer a definition of a mechanism as "a structure performing a function in virtue of its component parts, component operations, and their organisation." Mechanistic explanation proceeds by identifying these parts and identifying how their organised activity gives rise to the phenomenon of interest. Unlike the deductive-nomological framework, which seeks to explain higher-level phenomena by reducing them to lower-level laws, mechanistic explanation is explicitly *multilevel*. That is, it seeks to explain how phenomena at one level are *caused* by the operation of organised parts at the level below (Craver 2007). Multilevel mechanistic reduction therefore rejects both traditional reductionism, which seeks to replace theories at the higher level with theories at the lower level, as well as the strict autonomy of the special sciences (Bechtel 2008).

Mechanistic explanations are common across the life sciences. What distinguishes mechanistic explanation in the *cognitive sciences* is that the relevant mechanisms contain parts that *carry information,* such that their operations are controlled or determined by this information-carrying function (Bechtel 2008). That is, cognitive mechanisms involve *representations*.

Whilst this view is widely accepted, it gives rise to two deep challenges. First, what *distinguishes* these two kinds of mechanisms—that is, ordinary non-representational physical mechanisms and distinctively *information-processing* mechanisms? Second, how *can* the representational properties of structures within a mechanism be causally relevant to the capacities it exhibits? These questions are *related* insofar as an answer to the second plausibly answers the first: information-processing mechanisms are distinguished from other systems because the representational properties of their component parts are causally relevant to what they *do*. The questions are *difficult* because it is not obvious *how* this could be true. Representation is a *relational* property between internal states of a mechanism and some spatially or temporally distal (or even abstract or non-existent) target. Notoriously, it is difficult to explain how this relational property could be relevant to the systemic role of the internal vehicle that stands in this relation (cf. O'Brien 2015a).

To see the difficulties here, recall the influential idea about information-processing prominent in TCS. On this view, representations are *formal* entities whose systemic role is a function of their *syntactic* properties.[4] Specifically, one maps syntactic properties and inferential transitions onto physical states and state transitions, which in turn mimic semantic relations between such symbols under interpretation. As many have noted, this seems to make *representational status* of such symbols irrelevant to the mechanism's functioning, thereby leaving both of our questions unanswered (Searle 1980; Stich 1983).

Traditionally, there have been three prominent responses to this problem. The first accepts that there is no semantic difference between cognitive and non-cognitive systems in TCS (Searle 1980; Stich 1983). The second locates the difference at the level of *explanation*, not ontology, such that the semantic interpretation of physical symbol systems provides what Dennett (1987, p. 350) calls a "heuristic overlay" not reflective of its *real* functioning (cf. Egan 2013). Finally, the third contends that distinctively cognitive mechanisms *acquire* semantic properties through their causal interactions with the environment (Fodor 1987).

None of these responses is satisfactory from the perspective of cognitive neuroscience, however. The first effectively abandons its commitment to the explanatory importance of internal representations (Stich 1983). The second drives a wedge between ontological and explanatory questions that violates the basic principles of mechanistic explanation: if internal parts are not literally representational, a description of the mechanism should reflect this fact, not obscure it (Ramsey 2007). Finally, the third explains representational properties *in terms of the mechanism's functioning*, and thus violates the principle that internal representations *explain* such functioning (Churchland 2012).

We contend that *iconic* representations offer a solution to this problem. Iconic representations, recall, represent through resemblance to their target. As such, to explain how the representational properties of such representations could be causally relevant to a mechanism's functioning, we just need to show how their *resemblance* to some target domain could be *exploited* by the mechanism of which they are a part (Shea 2014; see Sect. 3.1 above). In other words, we need to show how the capacities a mechanism exhibits could be *causally dependent* on the degree to which its internal representations *resemble* the structure of their target. And—as Gładziejewski and Miłkowski (2017) point out—it is not difficult to see how this might come about. Because a *cognitive* mechanism's functioning is causally dependent on its ability to coordinate its behaviour—or the behaviour of the larger system of which it is a part—with some distal domain, the *resemblance* between its internal representations and the *structure* of that domain can be *causally relevant* to its success. For example, a rat's ability to navigate its ambient environment is causally dependent on the degree to which its hippocampal map of that environment accurately recapitulates its spatial structure (Gładziejewski and Miłkowski 2017). A mechanism that relies upon iconic representation is thus causally indebted to the accuracy of its internal representations—subject,

[4] Or, more accurately, the physical properties these syntactic properties are systematically mapped onto.

of course, to the qualification that *maximum* accuracy is neither feasible nor attractive (see Sect. 3.1 above).

Given this, iconic representations are *sufficient* for mechanistic explanation of distinctively information-processing systems, and offer a neat explanation of how the representational properties of internal vehicles can be causally relevant to a mechanism's functioning. We think one can plausibly go further, however: iconic representation is not just sufficient but *necessary* for this kind of explanation (O'Brien 2015a; O'Brien and Opie 2004).

To see why, note that a genuinely representational explanation of some phenomenon must satisfy the following condition: the properties by which internal representations *represent* must be responsible for their systemic role within the broader mechanism of which they are a part. Without this, the vehicles' representational properties become redundant. For example, suppose we simply *decide* to interpret the mechanism responsible for a bird's navigational abilities in terms of a conversation between Churchill and Hitler. In this case, the properties by which the mechanism components represent—namely, pure *convention* and *interpretation*—are completely irrelevant to their systemic role. Representation thus becomes a matter of observer-dependent overlay.

With iconic representations, by contrast, the property responsible for their representation of a given domain—namely, resemblance—doubles up as the property by which they perform their systemic role. The reason this is *possible* is that both properties are mediated by the representation's *structure*, which can simultaneously *re-present* or *stand in for* characteristics of the target *and* be causally relevant to a mechanism's broader functioning (O'Brien 2015a). If one turns to the sole serious *alternative* to resemblance-based accounts of representation, however, namely *causal* or *indicator*-based accounts, they fail this desideratum: the properties by which vehicles represent in such accounts—namely, the presence of some causal relationship between internal and external states—cannot *itself* be relevant to their systemic role (Bechtel 2008). To see this, consider the case of a thermostat. The mere fact that internal states are *caused* by external states cannot explain the systemic roles of the former in the broader regulative mechanism. For this reason, this causal relationship in thermostats *mediates* a relation that *can* be exploited by the broader mechanism: namely, the correspondence or *resemblance* relation between the height of mercury in the thermometer and the ambient room temperature (O'Brien 2015a).

For this reason, we suspect that iconic representations are not just sufficient but positively necessary for genuinely mechanistic explanations in which *representational properties* play a systemic role. Before turning to examples of such explanations in cognitive neuroscience, we briefly dispel a possible misinterpretation of our argument.

Our thesis is this: *cognitive neuroscience both entails (a prior) and exhibits (a posteriori) an iconic approach to cognitive representation*. Crucially, our thesis is emphatically *not* that such iconic representations are *unique* to cognitive neuroscience and thus *absent* from what is ordinarily thought of as TCS. This would be patently false. Indeed, Ramsey (2007) has argued persuasively that a central construct in TCS is precisely the "S-representation" (see also Cummins 1989). As a claim about how the concept of internal representation was *initially* understood in TCS, we think Ramsey's claim is simply mistaken. Indeed, when Johnson-Laird (1983) advanced his model-

based, iconic account of cognitive representation in the 1980s, he explicitly positioned it in *opposition* to the dominant understanding of representation that had preceded it in TCS—which, according to Johnson-Laird (1983 x), had neglected "a crucial issue: what it is that makes a mental entity a representation *of* something." Further, as Ramsey (2007) himself notes, the vast bulk of work in "naturalistic psychosemantics" that attempts to explain *how* internal representations posited by cognitive science acquire their contents explicitly shuns resemblance-based accounts in favour of causal or informational approaches to content determination (cf. Fodor 1987; Hutto and Satne 2015).

Nevertheless, dissatisfaction with the early programme of TCS and with prevalent confusions around the concept of internal presentation *did* motivate numerous authors within that tradition to embrace effectively iconic understandings of cognitive representation (e.g. Gallistel 1993; Johnson-Laird 1983). Many researchers typically thought to work within the "classical" programme have thus provided some of the most articulate defences of our foregoing argument that genuinely representational explanations of a system's behaviour require iconic representations (cf. Isaac 2013). Our claim is that cognitive neuroscience is committed to a resemblance-based approach to cognitive representations, not that this resemblance-based approach is *unique* to cognitive neuroscience. The latter claim would be absurd: after all, *Aristotle* defended an iconic approach to mental representation. Our title includes the "*return* of resemblance" for a reason.

4.2 Cognitive neuroscience and iconic representations

If our analysis is correct, and *if* cognitive neuroscience advances mechanistic explanations of genuinely representational systems, we should expect to find such iconic forms of representation within information-processing models in cognitive neuroscience. Our contention is that this is just what one does find if one examines research in cognitive neuroscience. Other philosophers have recently conducted excellent work on just this topic (cf. Churchland 2012; Gładziejewski and Miłkowski 2017; Kiefer and Hohwy 2017; Ryder 2004; Shea 2014; Williams 2017). Any exhaustive overview of such research—or of research into cognitive representation more generally in cognitive neuroscience—is impossible in a paper of this scope. Instead, in this section we focus on two examples that we think are especially important: first, the emergence of predictive models that replicate the dynamical structure of target domains to facilitate various cognitive functions (Sect. 4.2); and second, the recent emergence of powerful tools for uncovering the neural encoding of categorical knowledge (Sect. 4.4).

4.3 Dynamical models

Begin with the cerebellum. The histological structure of the cerebellum has been known for well over a century since the pioneering work of Santiago Ramón y Cajal (see Ramón y Cajal 1989) and it may be for this reason that the cerebellum was the target for those working in computational neuroscience while the discipline was still in its infancy (e.g., Marr 1969).

The microstructure of the neural connections within the cerebellum is very well understood in part because of the very regular pattern of organisation where a number of different cell types are organised in a stereotyped pattern (see Llinás 1975). This pattern resembles something like a grid with parallel fibers forming synapses onto the dendrites of Purkinje cells. Pellionisz and Llinás (1979) proposed that this structure might be ideally suited to perform a particular kind of computational operation, namely coordinate transformation. Mathematically, coordinate transformation occurs by multiplying an m-dimensional input vector by a m × n transformation matrix to yield an n-dimensional output vector. This model proposed that the grid-like arrangement of synapses in the cerebellum is able to act like the transformation matrix with the synaptic weights providing the values and the firing frequencies of the input and output cells acting as the input and output vectors. The input and output vectors are functions of time—or more specifically time-frequency functions—allowing them to, for example, dynamically track the movement of limbs. Moreover, the activity of the output cells in this arrangement can be mathematically interpreted as representing a Taylor expansion, and while the mathematics here are not of vital importance, a crucial property emerges when they are considered in this way. That is, these functions can act as *predictions*. Indeed, the notion that the cerebellum implements a predictive model has been foundational in neurocomputational approaches to action control (see Wolpert 1997).

This approach to action control makes extensive use of predictive models—in particular, models of the forward dynamics and inverse dynamics of the musculoskeletal systems (Wolpert 1997, 1998). The model of the inverse dynamics is used as a planner because it transforms a desired goal state into the motor commands required to bring about that goal—that is, it computes the motor commands necessary to achieve the goal. A model of the forward dynamics can be used to predict how the body will move in response to a particular set of motor commands or the sensory consequences that this movement will generate. Focusing just on forward models (for the sake of brevity), we can see that these predictions play a crucial role in the control of action. Forward models allow the motor system to engage in feedback control of action in the presence of sensory delays and noise. Without being able to make use of feedback during the performance of an action it would not be possible to update an action plan after it has been initiated. But feedback control allows the system to correct for any deviance in the planned trajectory by comparing the actual position with the predicted position. In the case of the action control system, however, sensory signals take too long to travel from the periphery to central areas to be useful and, furthermore, they can be contaminated with noise. In this case, predictions about the limb's position, generated by the forward model, can be used to stand in for actual sensory feedback and motor plans can be updated on the basis of this virtual sensory feedback or, to mitigate the effects of sensory noise, virtual feedback and actual feedback can be combined, with each source weighted according to their expected precision.

These forward predictive models of the body's dynamics are an example of a broader class of models, known as emulators (Grush 1997). The notion of emulators derives from control theory, the branch of engineering that seeks to understand how to control a target system (the plant) so as to bring about a desired behaviour. The emulator models the plant by replicating its dynamics and, therefore, replicating how the plant would

behave in response to certain control commands. As such, emulators function as iconic representations. By capturing the temporal evolution of the target system, the different states of the emulator would stand in relation to each other in the same way that those states stand in relation to each other in the target system—that is, it would represent by resemblance. And importantly, the emulator can be used offline so that control commands are issued to the emulator alone without producing any behavior in the plant. The system, equipped with an emulator like this, is now able to try out commands without producing behavior. The system can plan, it can entertain counterfactuals, and it can engage in other kinds of behavior that we usually think representational systems are capable of (Grush 2004). For these reasons, such emulators can play functional roles that merely *detector-based* systems cannot.

While models of action control may seem a long way off from the kind of phenomena that TCS was interested in, emulator models and similar iconic representations have been pushed far beyond the domain of action control. Some examples include language and dialog (Pickering and Garrod 2004), the sense of agency (Frith et al. 2000), the sense of embodiment (Carruthers 2013), perceptual prediction (Colling et al. 2014b; Wilson and Knoblich 2005), and accounts of joint planning and collective intentions (Butterfill 2015; Colling 2017), which serve as alternatives to the propositional accounts of Searle (2002) and Bratman (1993). In addition, while the work of Pellionisz and Llinás on the computations performed by the cerebellum we outlined above can be seen as foreshadowing the emulation theory of representation, the emulation theory of representation can itself be seen as foreshadowing *predictive processing*, an extremely influential emerging framework on brain function to which we return in Sect. 5.2.

Further, it has been argued that even cognitive representations usually understood in the form of "detector"-style representations are better understood within the broader framework of predictive models we have outlined here. For example, the extremely influential research on "mirror neurons" in psychology and neuroscience—namely, neurons that are active not just when an organism performs an action but when it observes a conspecific performing the same or similar action—are often understood through the framework of *motor resonance* (e.g. Gallese and Goldman 1998; cf. Pellegrino et al. 1992). Nevertheless, many have noted that this effectively detector-based framework seems to be inconsistent with the informational and functional profile of such neurons (cf. Csibra 2008; Umiltà et al. 2001), which are inherently *predictive* in character and structured into broader networks from which they derive their functional roles, leading some to rename them *emulator neurons* (Csibra 2008) and situate them which the iconic account of representation we outline here (Colling et al. 2013, 2014b; cf. also Clark 2016 ch. 7 and Kilner et al. 2007).

4.4 Resemblance and cognitive neuroimaging

One of the primary tools that cognitive neuroscience has made use of for uncovering neural mechanisms has been functional neuroimaging—techniques such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG). These techniques attempt to isolate particular brain regions

involved in particular cognitive capacities or to identify the timing of particular subcomponents of a larger cognitive capacity. Therefore, they can—or at least should (Colling and Robers 2010; McCauley and Bechtel 2001)—represent the first step in the mechanistic decomposition of a cognitive capacity. Until recently, these techniques have not been suited to uncovering representations proper but have rather been restricted to detecting neural signals that might correlate with some stimulus or cognitive task. This, however, has changed with the development of newer techniques that aim to uncover the structure and content of representations more directly. These techniques, which are based on neural decoding methods, endorse a resemblance-based, or iconic, account of cognitive representation. Neural decoding analyses exploit the idea that if the brain encodes information in patterns of neural activity then it should be possible, using the right techniques, to decode this information and uncover the content in a particular pattern of neural activity. In a typical neural decoding experiment, the experimenter might, for example, show the participant images from two different categories while recording brain activity using fMRI, EEG, or MEG (Grootswagers et al. 2017; Ritchie et al. 2017). A linear classifier is then trained to distinguish the patterns of activity produced in response to one class of stimuli from the patterns of activity produced in response to a different class of stimuli. In effect, the brain activity in response to each stimulus is treated as a point in high-dimensional activation space, and the classifier works by trying to place a decision boundary between the points corresponding to the two classes of stimuli.

Neural decoding methods themselves do not make any strong claims about the nature of cognitive representation. Specifically, a decoder could successfully distinguish the brain activity produced in response to stimulus A without that brain activity representing 'A' unless that statistical regularity was actually exploited by the brain (Ritchie et al. 2017). However, a related technique, known as Representational Similarity Analysis (RSA; Kriegeskorte and Kievit 2013), which employs neural decoding, plausibly *can* support claims about the nature of mental representation.

RSA attempts to uncover the *structure* of neural activation spaces. Rather than simply trying to distinguish the brain activity in response to stimulus A from the brain activity in response to stimulus B, an RSA analysis asks about the *pattern of relationships* between classes. For example, if a participant were shown images of different kinds of animals, for example, pigs, sheep, cows, dogs and cats, RSA could test how similar the pattern of neural activity is for each pairwise combination of stimuli. From this, it might be possible to discern whether there is a greater degree of similarity in the neural activity in response to sheep, cows, and pigs than there is between the neural activity to cows and cats, and this might create clusters of neural similarity that might correspond to our semantic categories of pets and farm animals (see Contini et al. 2017; Wardle et al. 2016; Ritchie and Carlson 2016 for examples of how RSA has been used, and see Ritchie et al. 2017 for a philosophical introduction to RSA).

Importantly, however, RSA can go further than simply uncovering semantic categories. RSA can also be used to uncover the *structure* of semantic categories. We know from philosophical work (Wittgenstein 1953) and work in cognitive psychology (Rosch and Mervis 1975) that semantic categories aren't formed by each exemplar within the category sharing an attribute with every other item in the category. Rather, each exemplar might share one attribute with some, but not all, other exemplars. As a

result, the exemplars in the category might be linked by a family resemblance relation or a web of overlapping attributes. A consequence of this is that some exemplars within the category might be more *prototypical* of the category than others. The structure of the semantic category allows us to see, for example, that football is more prototypical of the category 'sport' than golf. Or that football is more similar to rugby on the dimension of 'sport'-ness than football is to golf. These are claims about the similarity relations between the content of representations. These structures have been probed in psychological experiments using reaction time measures, with the general finding being that more prototypical exemplars are categorised into the appropriate category quicker. However, in the case of iconic representation, we should expect to find resemblance relations at the level of content mirrored at the level of the vehicle (consider how two maps would be similar if they were maps of two geographically similar cities). Therefore, it might be possible to ask the question: is the pattern of relationships between tokens in the represented domain mirrored at the level of the brain? That is, can we examine the brain directly and find evidence for the second-order resemblance relation advanced by proponents of iconic representation?

This is indeed what was done by Richie, Carlson, and colleagues (Carlson et al. 2014; Ritchie et al. 2015; see also Ritchie and Carlson 2016) by using RSA to generate neural activation spaces for images of animate and inanimate objects. Each token in these two categories was represented as a point in this activation space. A decision boundary, which maximally separated these two classes, was then drawn through this activation space. With the decision boundary in place, the distance between each token and the decision boundary could be measured, and this distance could be correlated with reaction times on an animate/inanimate judgment task. The results showed that the distance from the decision boundary predicted the reaction times—that is, the closer an exemplar was to the decision boundary separating animate and inanimate objects the slower people were to categorize it as animate. As reaction times have been shown to mirror the semantic relations between tokens in a category (Rosch and Mervis 1975), these experiments provide evidence that these relations are also mirrored at the level of neural population codes, exactly as one would expect if the brain made use of iconic forms of representation.

4.5 Summary

In this section we have argued on *a priori* grounds that mechanistic explanations of information-processing systems requires iconic representations—a lesson we have sought to illustrate by focusing on two especially promising areas of research in actual cognitive neuroscientific practice. This includes not only work that posits predictive dynamical models, but also recent work in cognitive neuroimaging that is explicitly directed at uncovering the structure of iconic representations in brain.

5 Prediction, regulation, and the model-making cortex

Boone and Piccinini (2015) rightly chastise TCS for being insufficiently attentive to mechanistic considerations in cognitive theorizing. Specifically, they argue that the

methodological assumptions of TCS are insensitive to the important ways in which "structures *constrain* functions and vice versa" (Boone and Piccinini 2015, p. 1522, our emphasis). We are in full agreement with this thesis, and have argued that a mechanistic approach to the explanation of information-processing in cognitive systems mandates an iconic account of cognitive representation.

However, we contend that there has also been an underappreciated shift in the way in which cognitive neuroscience addresses *functional* questions *independently of structural constraints*. That is, there has been a dramatic shift in the way in which cognitive neuroscience identifies the *problems* that cognitive systems (or sub-systems) solve.[5] This shift has been driven not *just* by greater sensitivity to the constraints imposed by the structure of neural mechanisms, but by a better appreciation of *functional constraints* that TCS was largely oblivious to. In this section, we note two of the most exciting trends in this respect—the emergence of broadly *global* theories of brain function that focus on *regulation* and *prediction* (Sect. 5.1)—and we explain how this shift has been equally important in driving cognitive neuroscience towards iconic representations (Sect. 5.2).

5.1 Homeostatic prediction machines

Sterling and Laughlin (2015 xvi) open a recent textbook with the claim that "the core task of all brains…. is to regulate the organism's internal milieu." In line with much recent work (Anderson 2014; Barrett 2017; Cisek 1999; Friston 2010; Hohwy 2013; Seth 2015), they argue that this overarching regulative lens on brain function can be exploited to derive general "principles of neural design" that might guide the very kinds of mechanistic explanation Boone and Piccinini note as central to cognitive neuroscience. As Barrett (2017, p. 3) puts it: "a brain did not evolve for rationality, happiness or accurate perception. All brains accomplish the same core task: to efficiently ensure resources for physiological systems within an animal's body (i.e. its internal milieu) so that an animal can grow, survive and reproduce."

This regulative perspective on brain function has deep roots in traditions like mid-twentieth century cybernetics (Conant and Ashby 1970), perceptual control theory (Powers 1973), dynamical systems theory-approaches to cognition (Van Gelder 1995), and a recent wave of work in theoretical neuroscience and biology that focuses on homeostasis and allostasis (the active process of achieving homeostasis) as overarching principles of brain function (for a representative sample, see: Anderson 2014; Barrett 2017; Corcoran and Hohwy 2018; Friston 2009, 2010; Seth 2015; Sterling and Laughlin 2015). Indeed, we saw it above in Grush's (1997; 2004) important work on emulators (Sect. 4.2), in which general characteristics of cognitive representations are derived from a control-theoretic perspective on brain function (see also Bechtel 2008, 2009).

At its deepest, the central idea underlying this efflorescence of work is that biological systems are distinctive in acting upon their environments to maintain their structural integrity and the homeostasis of their essential variables in a way that keeps

[5] These are questions that arise at what Marr (1982) called the "computational" level of explanation.

them far from thermodynamic equilibrium (Bechtel 2009; Friston 2010). In other words, they actively self-organize around their homeostatic set-points and thus somehow avoid the general tendency towards increasing disorder described by the second law of thermodynamics (Anderson 2014; Friston 2009, 2010). Brains are the control systems principally responsible for facilitating this process of self-organization in the organisms of which they are a part, mitigating the impact of environmental changes that push the organism outside its optimal states. For many, this important observation can be used to extract a foundational "job description" for brains: "to exert control over the organism's state within its environment" (Cisek 1999, pp. 8–9) and thus "maintain organism-relevant variables within some desired range" (Anderson 2015, p. 7).

Advocates of this regulative perspective contend that it has profound implications for our understanding of cognition (Cisek 1999; Friston 2009, 2010; Sterling and Laughlin 2015). Before turning to one manifestation of this recent stance, we note how radically different this approach to *functional* considerations is to that which characterised TCS: rather than functional decomposition of cognitive sub-systems in terms of analyses of the problems they solve that are largely unconstrained either structurally *or functionally* by lower-level sciences, this regulative perspective on brain function draws on theoretical considerations from statistical physics, theoretical biology, and mathematical and engineering frameworks on the nature of control and self-organization (see below) (cf. Bechtel 2008, pp. 159–201).

Alongside this turn to regulation, a similarly dramatic shift in recent years has been the turn towards *prediction* and *anticipatory* neural dynamics as a central variable in adaptive success (Barr 2011; Clark 2016; Downing 2009; Hohwy 2013). Downing (2009, p. 39), for example, notes that prediction often went "unappreciated as a fundamental component of intelligence" in TCS, in contrast to a widespread contemporary recognition that it "represents a fundamental principle of brain functioning which is at the core of cognition." Barr (2011 v) concurs, reporting a broad consensus that prediction is "a universal principle in the operation of the human brain." "The term "predictive brain", Bubic et al. (2010, p. 2) argue, "depicts one of the most relevant concepts in cognitive neuroscience."

What has induced this turn towards prediction as a fundamental operating principle of cognition? In part, it has been justified by recourse to very general considerations concerning adaptive success (Ryder 2004), which are in turn often linked to the foregoing control-theoretic perspective (see Sect. 5.2 below) (see Friston 2009, 2010; Sterling and Laughlin 2015). For example, any adaptive system that is not purely *reactive* must be able to *anticipate* changes in its environment, and *any* form of effective *intervention* in one's environment mandates sufficiently accurate information about its likely effects. As Llinás (2001, p. 21) puts it, "the capacity to predict the outcome of future events—critical to successful movement—is, most likely, the ultimate and most common of all global brain functions." Further, the turn to prediction has been partially taken over from work in machine learning, where prediction-driven learning has been central in explaining how artificial neural networks can exploit the statistics of their sensory input to learn about the structure of its *source* without guided supervision (Kiefer and Hohwy 2017).

In addition, purely reactive systems are extremely energetically inefficient (Sterling and Laughlin 2015): among other things, anticipatory dynamics enable brains to focus

only on functionally significant *surprises* in sensory input, effectively ignoring much the incoming signal (see below). More specifically, prediction has played a pivotal role in explaining (among numerous other phenomena) how brains: (i) disambiguate noisy and ambiguous sensory inputs (Hohwy 2013); (ii) disambiguate self-induced sensory inputs from environmentally induced ones (Blakemore et al. 2000); (iii) overcome various time-delays in incoming sensory signals in effective sensorimotor processing (see Sect. 4.2) (Franklin and Wolpert 2011); (iv) simulate "off-line" environmental processes (Barr 2011); and much more (for excellent overviews, see Barr 2011; Bubic et al. 2010; Clark 2016). This importance of prediction to so many aspects of cognitive functioning motivates Hawkins and Blakeslee (2004, p. 89) to declare that "prediction is not just one of the things your brain does. It is the *primary function* of the neocortex, and the foundation of intelligence. The cortex is an organ of prediction."

Recently, this emphasis on regulation and prediction has reached its apotheosis in "predictive processing" (Clark 2016; Friston 2010; Hohwy 2013), a global theory of brain function which comprises one manifestation of a broader trend of theoretical approaches that view the mammalian cortex as a general-purpose model-making mechanism whose chief function is *prediction* (e.g., Friston 2009, 2010; Hawkins and Blakeslee 2004; Ryder 2004). *Predictive processing* is by far the most ambitious and influential of such approaches, touted by some as the "first unified theory of the brain" (Huang 2008)—a "paradigm shift" (Friston et al. 2017, p. 1) that constitutes "the most complete framework to date for explaining perception, cognition, and action in terms of fundamental theoretical principles and neurocognitive architectures" (Seth 2015, p. 1). Roughly, it states that brains self-organize around a single, overarching imperative: the *minimization of long-term prediction error*, the mismatch between the sensory data they predict on the basis of their model of the world and the sensory data they receive from that world (Clark 2016; Friston 2010; Hohwy 2013). When recapitulated up the hierarchical structure of the neocortex, this process of prediction error minimization is supposed to account for "perception and action and *everything mental in between*" (Hohwy 2013, 1, our emphasis].

We don't propose to take a stand on the *truth* of predictive processing. An enormous amount has been and continues to be written about it, both in the scientific and philosophical literature (see Clark 2016 and Hohwy 2013 for excellent overviews, and Colombo and Wright 2017 for criticism). It would be impossible to provide a full overview—let alone critical evaluation—of the theory in the space here. Further, as many have noted, predictive processing is at best a *mechanism sketch* (cf. Kaplan and Craver 2011b) at present, pitched largely at Marr's (1982) "algorithmic" level of description with only tentative proposals about how the functional roles it identifies are realised in cortical circuitry (cf. Brodski et al. 2015; Gordon et al. 2017; Weilnhammer et al. 2017).[6] Nevertheless, it nicely illustrates the points we are making, and manifests how a regulative and thoroughly predictive perspective on brain function harmonises with an iconic approach to cognitive representation.

First, predictive processing situates brain function within the context of homeostatic regulation as outlined above. Specifically, it contends that prediction error

[6] We thank an anonymous reviewer for pressing this point.

minimization is a special case of a more fundamental imperative in biological systems to self-organize under conditions tending towards increasing disorder, whereby the brain's ceaseless efforts to align top-down predictions with the incoming signal maintain the broader organism within those biophysical states consistent with homeostasis (Friston 2009, 2010). For this reason, Yufik and Friston (2016) describe its functional approach as a "physics of the mind," an attempt to view cognition from the perspective of statistical physics and theoretical biology.

Second, predictive processing unsurprisingly places *prediction* at the functional core of cognition. It is only by minimizing the error in their predictions of the incoming signal that brains can maintain the organisms of which they are a part within their optimal states and thus fulfil their homeostatic function (Friston 2010; Friston et al. 2017). Crucially, to effectively minimize the error in their predictions of the incoming *signal* requires that they effectively predict how the environmental *causes* of such signals are likely to evolve under various kinds of alteration and intervention (Clark 2016, p. 6). That is, a brain can effectively predict the incoming signal only by effectively predicting the worldly causes of that signal and how they are likely to alter and evolve.

Third, and most importantly for our purposes, at the centre of predictive processing is a commitment to a thoroughly *iconic* account of cognitive representation, whereby cortical dynamics come to recapitulate regularity structures in the body and environment and thereby realise a rich, hierarchically structured *generative model* with which to anticipate the incoming signal (Gładziejewski 2015; Kiefer and Hohwy 2017; Williams 2017). Specifically, it is only by inducing and updating a hierarchically structured dynamical model of the bodily and environmental causes of their sensory input that predictive brains can fulfil their regulative function (Seth and Friston 2016). In this way "neuroanatomy and neurophysiology can be regarded as a distillation of statistical or causal structure in the environment disclosed by sensory samples" (Seth and Friston 2016, p. 3), and the brain's generative model "inherits the dynamics of the environment and can predict its sensory products accurately" (Kiebel et al. 2009, p. 7). The generative models in this framework are thus effectively versions of the dynamical (emulator) models introduced in Sect. 4, albeit ones that encode *probabilistic* dependencies between environmental variables rather than deterministic ones, in a way familiar from Bayesian networks in machine learning (Pearl 2009; cf. Kiefer and Hohwy 2017).

5.2 Regulation, prediction, resemblance

Crucially, this commitment to iconic representation is not *incidental* in predictive processing. Specifically, iconic representations are positively *mandated* by its commitment to a thoroughly regulative and predictive conception of brain function.

To see this, note that the free-energy principle effectively advances an information-theoretic vindication of the "good regulator theorem" advanced by Conant and Ashby (Conant and Ashby 1970; cf. Ashby 1952, 1956; and Seth 2015 for an excellent summary), according to which optimal regulation mandates the construction of a *model* that replicates the functionally significant structure of the system being regulated. Roughly, that is, the theorem asserts that any system whose function is to regulate

another system must—insofar as it is optimal—exploit a *stand-in* or *model* of that system that is isomorphic (i.e. structurally similar) to it. If this is right, and *if* the core task of all brains is homeostatic regulation, we should expect the fundamental kind of cognitive representations to be iconic representation—just as predictive processing asserts. As Conant and Ashby (1970, p. 89) famously put it:

> The theorem has the interesting corollary that the living brain, so far as it is to be successful and efficient as a regulator for survival must proceed, in learning, by the formation of a model (or models) of its environment.

What about prediction?[7] As noted above, predictive processing capitalizes on research in machine learning in which unsupervised learning is driven by the exploitation of a generative model of environmental causes (Kiefer and Hohwy 2017). The punchline of this research is that the only way in which such a system can successfully *predict* or *anticipate* the sensory input generated by the world is by effectively *becoming that world*—that is, by *inverting* and thereby *recapitulating* the process by which sensory input is generated in its structure and dynamics (Kiefer and Hohwy 2017; Williams 2017). The link between *prediction* and *iconic representation* in this context is thus not incidental. What Clark (2016, p. 299) calls the "prediction-and-generative-model-based" approach to the mind is fundamentally committed to iconic representations in which the structure of internal representations in the brain come to replicate the structure of the generative process by which sensory input impinges upon it (Gładziejewski 2015).

Predictive processing is thus a concrete illustration of the deep theoretical links between regulation, prediction, and iconic representation. Nevertheless, as noted above, these links are in principle independent of predictive processing's theoretical idiosyncrasies, which is still in its infancy, and which is yet to develop the fully mechanistic details required for full maturity. What it nicely reveals, however, is that the shift towards an iconic brain is driven as much by a greater appreciation of purely *functional* considerations as it is by the demands of multilevel mechanistic explanation that Boone and Piccinini rightly stress.

6 Conclusion

We have argued in this paper that the cognitive neuroscience revolution described by Boone and Piccinini has engendered a dramatic shift away from traditional ways of thinking of cognitive representation and back towards a truly ancient one—towards resemblance-based representations. Specifically, we have argued that this shift has been driven both from the demands of mechanistic explanation and from a greater appreciation of the problems solved by bio-cognitive systems, and we have sought to illustrate this shift by reference to what we regard as some of the most important and exciting developments in recent neuroscience. We end this paper by noting three

[7] Craik (1943) argues for both the centrality of prediction to cognition and the necessity of internal structural models for prediction.

issues that have arisen in our discussion that deserve greater consideration than space constraints have allowed us to provide here.

First, we have tied iconic representation to exploitable structural similarity and to homeostatic regulation. For some, this kind of move implies an implausible "pan-representationalism" (Ramsey 2007) that trivialises the concept of cognitive representation beyond recognition. We disagree. "Pan-representationalism" *as such* cannot be an objection to a view. How widely a concept applies must be decided by empirical inquiry, not pre-theoretical intuitions, and we think the views defended in this paper highlight the important *scientific* discovery that the exploitation of internal models is characteristic of all adaptive systems, thereby highlighting a profound *continuity* between life and mind. Nevertheless, addressing these issues requires more space than we could allocate here.

Second, a general worry often levelled against a focus on iconic representations is that they will not *scale* to explain the more sophisticated cognitive capacities exhibited by organisms like ourselves. Whilst we regard this worry as misplaced, we recognise that a central challenge for the vision of an iconic brain that emerges in the current paper is how the information-processing strategies they employ might interact with—and become *augmented* and *transformed* by—the *public* combinatorial symbol systems that TCS transposed into the head.

Finally, our paper has necessarily focused only on a selective range of work in cognitive neuroscience, and has therefore neglected a good deal of contemporary research on the nature of cognitive representation and cortical information-processing. Perhaps the most exciting of this work is Eliasmith's (2013) recently developed "*semantic pointer*" approach to cognitive representation, which attempts to integrate the insights of classical symbolic approaches within a mechanistic framework. A more thorough exploration of our argument would show how this research—and other promising directions in cognitive neuroscience—could be integrated into the resemblance-based perspective on representation advocated here. We think it can be done, but doing it is a task for another paper.

Acknowledgements This work was supported by the Arts and Humanities Research Council. We would like to thank two anonymous reviewers for extremely helpful comments on an earlier version of this manuscript.

References

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

Anderson, M. L. (2015). Précis of after phrenology: Neural reuse and the interactive brain. *Behavioral and Brain Sciences*. doi:10.1017/s0140525x15000631.

Ashby, W. R. (1952). *Design for a brain*. London: Chapman and Hall.

Ashby, W. R. (1956). *An introduction to cybernetics*. London: Chapman and Hall.

Barr, M. (2011). *Predictions in the brain*. Oxford: Oxford University Press.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*. doi:10.1093/scan/nsw154.
Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science*, *22*(3), 295–318. doi:10.1207/s15516709cog2203_2.
Bechtel, W. (2008). *Mental mechanisms*. Hoboken, NJ: Taylor and Francis.
Bechtel, W. (2009). Constructing a philosophy of science of cognitive science. *Topics in Cognitive Science*, *1*(3), 548–569. doi:10.1111/j.1756-8765.2009.01039.x.
Bechtel, W. (2014). Investigating neural representations: The tale of place cells. *Synthese*, *193*(5), 1287–1321. doi:10.1007/s11229-014-0480-8.
Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441. doi:10.1016/j.shpsc.2005.03.010.
Blakemore, S., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself? *Neuroreport*, *11*(11), R11–R16. doi:10.1097/00001756-200008030-00002.
Boone, W., & Piccinini, G. (2015). The cognitive neuroscience revolution. *Synthese*, *193*(5), 1509–1534. doi:10.1007/s11229-015-0783-4.
Bratman, M. E. (1993). Shared intention. *Ethics*, *104*(1), 97–113. doi:10.1086/293577.
Brodski, A., Paasch, G.-F., Helbling, S., & Wibral, M. (2015). The faces of predictive coding. *Journal of Neuroscience*, *35*(24), 8997–9006. doi:10.1523/JNEUROSCI.1529-14.2015.
Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*,. doi:10.3389/fnhum.2010.00025.
Butterfill, S. A. (2015). Planning for collective agency. In C. Misselhorn (Ed.), *Collective agency and cooperation in natural and artificial systems* (pp. 149–168). Heidelberg: Springer.
Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience*, *26*(1), 132–142. doi:10.1162/jocn_a_00476.
Carruthers, G. (2013). Toward a cognitive model of the sense of embodiment in a (rubber) hand. *Journal of Consciousness Studies*, *20*(3–4), 33–60.
Chemero, A. (2001). Dynamical explanation and mental representations. *Trends in Cognitive Sciences*, *5*(4), 141–142. doi:10.1016/S1364-6613(00)01627-2.
Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: The MIT Press.
Chomsky, N. (1959). Review of "verbal behavior". *Language*, *35*(1), 26–58. doi:10.2307/411334.
Churchland, P. (2012). *Plato's camera*. Cambridge, MA: MIT Press.
Cisek, P. (1999). Beyond the computer metaphor: Behaviour as interaction. *Journal of Consciousness Studies*, *6*(11–12), 125–142.
Clark, A. (1997). *Being there*. Cambridge, MA: MIT Press.
Clark, A. (2016). *Surfing uncertainty*. Oxford: Oxford University Press.
Colling, L. J. (2017). Planning together and playing together. In M. Cappuccio (Ed.), *Handbook of embodied cognition and sport psychology*. Cambridge, MA: The MIT Press.
Colling, L. J., Knoblich, G., & Sebanz, N. (2013). How does "mirroring" support joint action? *Cortex*, *49*, 2964–2965. doi:10.1016/j.cortex.2013.06.006.
Colling, L. J., & Robers, R. P. (2010). Cognitive psychology does not reduce to neuroscience. In W. Christensen, E. Schier, & J. Sutton (Eds.), *ASCS09: Proceedings of the 9th conference of the Australasian society for cognitive science* (pp. 41–48). Sydney, Australia: Macquarie Centre for Cognitive Science.
Colling, L. J., Thompson, W. F., & Sutton, J. (2014). The effect of movement kinematics on predicting the timing of observed actions. *Experimental Brain Research*, *232*(4), 1193–1206. doi:10.1007/s00221-014-3836-x.
Colling, L. J., & Williamson, K. (2014). Entrainment and motor emulation approaches to joint action: Alternatives or complementary approaches? *Frontiers in Human Neuroscience*, *8*(26), 67. doi:10.3389/fnhum.2014.00754.
Colombo, M., & Wright, C. (2017). Explanatory pluralism: An unrewarding prediction error for free energy theorists. *Brain and Cognition*, *112*, 3–12. doi:10.1016/j.bandc.2016.02.003.
Conant, R., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97. doi:10.1080/00207727008920220.
Contini, E. W., Wardle, S. G., & Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia*,. doi:10.1016/j.neuropsychologia.2017.02.013.

Corcoran, A., & Hohwy, J. (2018). Allostasis, interoception, and the free energy principle: Feeling our way forward. In M. Tsakiris & H. De Preester (Eds.), *The interoceptive basis of the mind*. Oxford: Oxford University Press.
Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.
Crane, T. (2003). *The mechanical mind* (2nd ed.). London: Routledge.
Craver, C. F. (2007). *Explaining the brain*. Oxford: Oxford University Press.
Csibra, G. (2008). Action mirroring and action understanding: An alternative account. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Sensorimotor foundations of higher cognition* (pp. 435–459). Oxford: Oxford University Press.
Cummins, R. (1989). *Meaning and mental representation*. Cambridge, MA: MIT Press.
Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge, MA: MIT Press.
Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
Dennett, D. (2013). *Intuition pumps and other tools for thinking*. New York: Norton & Company.
di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, *91*(1), 176–180. doi:10.1007/BF00230027.
Downing, K. (2009). Predictive models in the brain. *Connection Science*, *21*(1), 39–74. doi:10.1080/09540090802610666.
Dretske, F. (1981). *Knowledge of the flow of information*. Cambridge, MA: MIT Press.
Egan, F. (2013). How to think about mental content. *Philosophical Studies*, *170*(1), 115–135. doi:10.1007/s11098-013-0172-0.
Eliasmith, C. (2013). *How to build a brain*. Oxford: Oxford University Press.
Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.
Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.
Fodor, J., & Pylyshyn, Z. (2015). *Minds without meanings: An essay on the content of concepts*. Cambridge, MA: MIT Press.
Franklin, D., & Wolpert, D. (2011). Computational mechanisms of sensorimotor control. *Neuron*, *72*(3), 425–442.
Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. doi:10.1016/j.tics.2009.04.005.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. doi:10.1038/nrn2787.
Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49. doi:10.1162/neco_a_00912.
Frith, C., Blakemore, S., & Wolpert, D. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *355*(1404), 1771–1788. doi:10.1098/rstb.2000.0734.
Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501. doi:10.1016/S1364-6613(98)01262-5.
Gallistel, C. (1993). *The organization of learning*. Cambridge, MA: The MIT Press.
Gallistel, C., & King, A. (2011). *Memory and the computational brain*. Hoboken, NJ: Wiley.
Gładziejewski, P. (2015). Predictive coding and representationalism. *Synthese*, *193*(2), 559–582. doi:10.1007/s11229-015-0762-9.
Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology & Philosophy*, *32*(3), 337–355. doi:10.1007/s10539-017-9562-6.
Goodman, N. (1969). *Languages of art*. London: Oxford University Press.
Gordon, N., Koenig-Robert, R., Tsuchiya, N., van Boxtel, J., & Hohwy, J. (2017). Neural markers of predictive coding under perceptual uncertainty revealed with hierarchical frequency tagging. *eLife*, *6*, e22749. doi:10.7554/eLife.22749.
Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. doi:10.1162/jocn_a_01068.
Grush, R. (1997). The architecture of representation. *Philosophical Psychology*, *10*(1), 5–23. doi:10.1080/09515089708573201.
Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*(3), 377–396. doi:10.1017/s0140525x04000093.
Haugeland, J. (1989). *Artificial intelligence*. Cambridge, MA: MIT Press.
Hawkins, J., & Blakeslee, S. (2004). *On intelligence*. New York: Henry Holt and Company.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175. doi:10.1086/286983.
Hempel, C. G., & Oppenheim, P. (1953). The logic of explanation. In H. Feigl & M. Brodbeck (Eds.), *Readings in the philosophy of science* (pp. 319–352). New York: Appleton-Century-Crofts, Inc.
Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.
Horst, S. (2016). *Cognitive pluralism*. Cambridge, MA: MIT Press.
Huang, G. (2008). Is this a unified theory of the brain? *New Scientist*, *2658*, 30–33.
Hutto, D., & Myin, E. (2013). *Radicalizing enactivism*. Cambridge, MA: MIT Press.
Hutto, D., & Satne, G. (2015). The natural origins of content. *Philosophia*, *43*(3), 521–536. doi:10.1007/s11406-015-9644-0.
Isaac, A. (2013). Objective similarity and mental representation. *Australasian Journal of Philosophy*, *91*(4), 683–704. doi:10.1080/00048402.2012.728233.
Johnson-Laird, P. (1983). *Mental models*. Cambridge, MA: Harvard University Press.
Kaplan, D. M., & Bechtel, W. (2011). Dynamical models: An alternative or complement to mechanistic explanations. *Topics in Cognitive Science*, *3*(2), 438–444. doi:10.1111/j.1756-8765.2011.01147.x.
Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627. doi:10.1086/661755.
Kiebel, S., Daunizeau, J., & Friston, K. (2009). Perception and hierarchical dynamics. *Frontiers in Neuroinformatics*,. doi:10.3389/neuro.11.020.2009.
Kiefer, A., & Hohwy, J. (2017). Content and misrepresentation in hierarchical generative models. *Synthese*. doi:10.1007/s11229-017-1435-7.
Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166. doi:10.1007/s10339-007-0170-2.
Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. doi:10.1016/j.tics.2013.06.007.
Llinás, R. R. (1975). The cortex of the cerebellum. *Scientific American*, *232*(1), 56–71. doi:10.1038/scientificamerican0175-56.
Llinás, R. R. (2001). *I of the vortex*. Cambridge, MA: MIT Press.
Marcus, G. (2003). *The algebraic mind*. Cambridge, MA: MIT Press.
Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology*, *202*(2), 437–470. doi:10.1113/jphysiol.1969.sp008820.
Marr, D. (1982). *Vision*. New York: Freeman.
McCauley, R. N., & Bechtel, W. (2001). Explanatory pluralism and heuristic identity theory. *Theory & Psychology*, *11*(6), 736–760. doi:10.1177/0959354301116002.
McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, *52*(1–2), 99–115. doi:10.1007/bf02459570.
Morgan, A. (2014). Representations gone mental. *Synthese*, *192*(2), 213–244. doi:10.1007/s11229-013-0328-7.
Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. doi:10.1145/360018.360022.
O'Brien, G. (2015). How does mind matter? Solving the content causation problem. In T. K. Metzinger & J. M. Windt (Eds.), *Open mind*. Frankfurt am Main: MIND Group. doi:10.15502/9783958570146.
O'Brien, G., & Opie, J. (2004). Notes towards a structuralist theory of mental representation. In H. Clapin, P. Staines, & P. Slezak (Eds.), *Representation in mind*. Amsterdam: Elsevier.
O'Brien, G., & Opie, J. (2015). Intentionality lite or analog content? *Philosophia*, *43*(3), 723–729. doi:10.1007/s11406-015-9623-5.
Pearl, J. (2009). *Causality* (2nd ed.). Cambridge: Cambridge University Press.
Peirce, C. S. (1931–1958). *Collected papers of Charles Sanders Peirce*. In: P. Hartshorne, P. Weiss, & A. Burks (Eds.) (Vols. 1–8). Cambridge, MA: Harvard University Press.
Pellionisz, A., & Llinás, R. R. (1979). Brain modeling by tensor network theory and computer simulation. the cerebellum: Distributed processor for predictive coordination. *Neuroscience*, *4*(3), 323–348. doi:10.1016/0306-4522(79)90097-6.
Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. doi:10.1017/S0140525X04000056.
Powers, W. T. (1973). *Behavior: the control of perception*. Hawthorne, NY: Aldine de Gruyter.
Ramón y Cajal, S. (1989). *Recollections of my life; translated by E. Horne Craigie*. Cambridge, MA: MIT Press.

Ramsey, W. (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.
Ritchie, J. B., & Carlson, T. A. (2016). Neural decoding and inner psychophysics: A distance-to-bound approach for linking mind, brain, and behaviour. *Frontiers in Neuroscience*, *10*(33), 310. doi:10.3389/fnins.2016.00190.
Ritchie, J. B., Kaplan, D. M., & Klein, C. (2017). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *British Journal for the Philosophy of Science*. doi:10.1093/bjps/axx023.
Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLOS Computational Biology*, *11*(6), e1004316. doi:10.1371/journal.pcbi.1004316.
Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. doi:10.1016/0010-0285(75)90024-9.
Ryder, D. (2004). SINBAD neurosemantics: A theory of mental representation. *Mind and Language*, *19*(2), 211–240. doi:10.1111/j.1468-0017.2004.00255.x.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. doi:10.1017/S0140525X00005756.
Searle, J. R. (2002). Collective intentions and actions. In *Consciousness and language* (pp. 90–105). Cambridge: Cambridge University Press.
Seth, A. K. (2015). The cybernetic Bayesian brain. In T. K. Metzinger & J. M. Windt (Eds.), *Open mind*. Frankfurt am Main: MIND Group. doi:10.15502/9783958570108.
Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160007. doi:10.1098/rstb.2016.0007.
Shea, N. (2014). VI—Exploitable isomorphism and structural representations. *Proceedings of the Aristotelian Society*, *114*(2pt2), 123–144. doi:10.1111/j.1467-9264.2014.00367.x.
Sterling, P., & Laughlin, S. (2015). *Principles of neural design*. Cambridge, MA: MIT Press.
Stich, S. (1983). *From folk psychology to cognitive science*. Cambridge, MA: MIT Press.
Turing, A. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2–42*(1), 230–265. doi:10.1112/plms/s2-42.1.230.
Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., et al. (2001). I know what you are doing: A neurophysiological study. *Neuron*, *31*(1), 155–165. doi:10.1016/S0896-6273(01)00337-3.
van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy*, *92*(7), 345–381. doi:10.2307/2941061.
Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S.-M., & Carlson, T. A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. *NeuroImage*, *132*, 59–70. doi:10.1016/j.neuroimage.2016.02.019.
Waskan, J. (2006). *Models and cognition*. Cambridge, MA: MIT Press.
Weilnhammer, V., Stuke, H., Hesselmann, G., Sterzer, P., & Schmack, K. (2017). A predictive coding account of bistable perception—A model-based fMRI study. *PLOS Computational Biology*, *13*(5), e1005536. doi:10.1371/journal.pcbi.1005536.
Williams, D. (2017). Predictive processing and the representation wars. *Minds and Machines*. doi:10.1007/s11023-017-9441-6.
Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, *131*(3), 460–473. doi:10.1037/0033-2909.131.3.460.
Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Wiley.
Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in Cognitive Sciences*, *1*(6), 209–216. doi:10.1016/S1364-6613(97)01070-X.
Wolpert, D. M. (1998). Multiple paired forward and inverse models for motor control. *Neural networks*, *11*(7–8), 1317–1329. doi:10.1016/S0893-6080(98)00066-5.
Yufik, Y. M., & Friston, K. J. (2016). Life and understanding: The origins of "understanding" in self-organizing nervous systems. *Frontiers in Systems Neuroscience*, *10*, 90. doi:10.3389/fnsys.2016.00098.